
Proceedings Paper:
Knill, K., Gales, M., Kyriakopoulos, K. et al. (4 more authors) (2018) Impact of ASR performance on free speaking language assessment. In: Interspeech 2018. Interspeech 2018, 02-06 Sep 2018, Hyderabad, India. International Speech Communication Association (ISCA) , pp. 1641-1645.

10.21437/interspeech.2018-1312

Interspeech 2018
2-6 September 2018, Hyderabad

Impact of ASR Performance on Free Speaking Language Assessment

K.M.Knill[1], *M.J.F.Gales*[1], *K. Kyriakopoulos*[1], *A. Malinin*[1], *A. Ragni*[1], *Y. Wang*[1], *A.P.Caines*[2]

[1]ALTA Institute / Engineering Department
[2]ALTA Institute / Computer Lab
Cambridge University, UK

{kate.knill,mjfg,kk492,am969,ar527,yw396}@eng.cam.ac.uk,apc38@cam.ac.uk

Abstract

In free speaking tests candidates respond in spontaneous speech to prompts. This form of test allows the spoken language proficiency of a non-native speaker of English to be assessed more fully than read aloud tests. As the candidate's responses are unscripted, transcription by automatic speech recognition (ASR) is essential for automated assessment. ASR will never be 100% accurate so any assessment system must seek to minimise and mitigate ASR errors. This paper considers the impact of ASR errors on the performance of free speaking test auto-marking systems. Firstly rich linguistically related features, based on part-of-speech tags from statistical parse trees, are investigated for assessment. Then, the impact of ASR errors on how well the system can detect whether a learner's answer is relevant to the question asked is evaluated. Finally, the impact that these errors may have on the ability of the system to provide detailed feedback to the learner is analysed. In particular, pronunciation and grammatical errors are considered as these are important in helping a learner to make progress. As feedback resulting from an ASR error would be highly confusing, an approach to mitigate this problem using confidence scores is also analysed.
Index Terms: speech recognition, spoken language assessment

1. Introduction

More than 1.5 billion people are predicted to be learning English as an additional language by 2020 [1]. Assessment of a learner's language proficiency is a key part of learning both in measuring progress made and for formal qualifications required e.g. for entrance to university or to obtain a job. It will be very difficult to train sufficient examiners for this many learners. Automatic systems for text and spoken language assessment have started to be deployed to assist. The speech systems mostly focus on reading aloud or tightly constrained tasks (e.g. [2, 3, 4]). A better reflection of a learner's ability to communicate orally is achieved through free speaking tasks where the learner is prompted to produce spontaneous speech. This is much harder to assess automatically due to the far greater diversity in such speech [5]. As can be seen in Figure 1, the ability of the automatic assesment system will be dependent on transcriptions produced by automatic speech recognition (ASR). This paper, therefore, considers how ASR perfomance impacts on free speaking language assessment. In addition, the ability to detect specific types of errors is investigated for future systems where learners can receive feedback on why they were awarded a specific grade, as has started to be deployed for text [6].

Both read aloud and free speaking tasks require the ASR to be capable of recognising non-native English speech. Across the Council of Europe CEFR levels [7] this varies from minimal (A1), through limited but effective (B1), to fully operational command of the spoken language (C2). The pronunciation, speaking rate, grammatical correctness, vocabulary and linguistic complexity are all affected by the first language (L1) of the speaker and by their level of proficiency. For free speaking tasks the ASR must also handle disfluent spontaneous speech with a wider, open vocabulary.

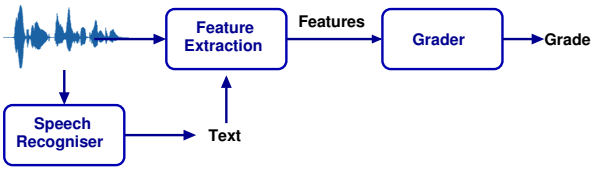


Figure 1: *Spoken language assessment auto-marker framework.*

A range of features are used within the auto-marker, primarily relating to audio and fluency characteristics. By using features derived from ASR (rather than manual) transcriptions in training of the auto-marker to match test conditions (Figure 1), the effects of ASR errors can be somewhat mitigated. Even with this mitigation, various papers have reported that improving ASR can lead to improvements in assessment with gains in machine-human correlation ranging from 0.02 to 0.7 e.g. [4, 8, 9]. Some features that may be beneficial for auto-marking are more affected by ASR errors, such as part-of-speech tags and features related to the spoken content, so their use has been limited at present [9, 10]. The ability to determine the content of a candidate's speech is also important to assess how relevant the candidate's response is to the question topic. Off-topic response detection [11, 12, 13] should, therefore, benefit from improvements in ASR transcription quality.

To help a learner make progress they need to receive detailed feedback on the errors they are making. At lower grades pronunciation errors dominate. In read aloud tasks the candidate responses can be compared to a matching reference response recorded by a native speaker. This is not possible for free speaking tasks so it relies on finding patterns based on the ASR transcripts. Higher grades need more feedback on grammatical errors. Again this relies on the ASR output to determine possible errors, made harder by the still open research questions of what is spoken grammar [14] or a "sentence" in spontaneous speech [15]. Any feedback must take into account the fact that the detected "error" may be the result of a mis-transcription by the ASR system rather than a true error made by the learner.

Section 2 presents the components in the auto-marking system including the ASR systems considered, the Gaussian Process grader used, and grader features investigated. Off-topic response detection is described in Section 3. Experimental re-

This research was funded under the ALTA Institute, Cambridge University. Thanks to Cambridge Assessment English for supporting this research and providing access to the BULATS data.

10.21437/Interspeech.2018-1312

sults on data from real exams are reported in Section 4. Finally conclusions are given in Section 5.

2. Auto-marking System

This section describes the auto-marking system components used to compare the impact of ASR peformance on assessment of free speaking by non-native learners of English. Data from non-native learners of English on the BULATS Business English exam [16] is used for training and test. The data was recorded in live exams with human examiners. It contains learners across the CEFR [7] proficiency levels, with the majority in the range A1-B2. The data contains a wide range of L1s.

2.1. ASR Systems

ASR performance is assessed in terms of word error rate (WER). For this paper two speaker independent ASR systems which have a significant difference on the test set are selected for investigation, arising from different training data sets and acoustic model (AM) architectures. The latter are close to and at state-of-the-art [1]

The first system, *System 1*, is a phonetic stacked hybrid DNN-HMM [10]. A bottleneck DNN is trained on the AMI corpus [17]. This corpus was selected as it contains high quality transcriptions of, mostly, non-native speakers of English. Filterbank features are used as input to the BN DNN. Tandem bottleneck (BN) and PLP features are input to a hybrid DNN-HMM AM, transformed by a HLDA transform [18]. The hybrid DNN output targets are global state-position context-dependent (triphone) targets [19] taken from a set of phonetic PLP GMM-HMMs trained on the same data. The AM is trained on 108 hours of recordings of Gujarati L1 candidates. Frame-weighted word level confidence scores are returned by the ASR engine [20] which have a piece-wise linear mapping applied for use in error detection. Pronunciations are taken from Combilex [21]. Out of lexicon words are pronunced using a Sequitur G2P system [22] trained on Combilex.

System 2 is a graphemic stacked hybrid DNN+LSTM-HMM joint decoding [23] system. All the neural nets are trained on approximately 334 hours of candidate recordings covering 28 L1s [2], including the 109 hours used for System 1. The hybrid DNN and LSTM output targets are global state-position tri-grapheme targets generated by a set of graphemic PLP GMM-HMMs trained on the same data. Word level confidence scores are returned from the Kaldi [24] decoder which are frame weighted and undergo a piece-wise mapping for use in error detection. The alphabet letters `/a-z/` form the base grapheme set, with two additional root graphemes, `/G00,G01/`, to model hesitation events. All words appear in the graphemic lexicon which includes attributes for apostrophes and partial words and boundary markers [10].

2.2. Grader Features

As for other auto-marker systems (e.g. [2, 25, 26]), the main input features to the grader are based on statistics derived from the speaker's audio and time aligned ASR hypotheses. The latter - at word, phone and consonant/vowel level - are used to derive proxies for speaker fluency, such as speaking rate, the mean duration of words and the fraction of disfluencies [27]. Audio features, such as the mean and standard deviation of the energy, are determined from the audio signal directly.

Lexical and grammatical features derived from statistical parses, such as part-of-speech (PoS) n-grams, can discriminate proficiency level on text data [28]. PoS tags are considered here. The PoS tags are derived from parse trees computed using RASP [29] on the ASR transcription after removing partial words and hesitations. Figure 2 illustrates parse trees generated from manual and ASR transcriptions of the same speech.

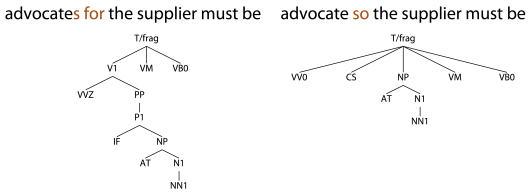


Figure 2: *Parse trees generated from manual (left) and ASR (right) transcriptions [27].*

2.3. Grader

The Gaussian Process grader proposed in [26] is used for the experiments in this paper. A GP is trained to predict the auto-marking score based on a set of input features as described above. The training data is labelled with scores given by the original human examiners. These are the true overall scores obtained by a candidate i.e. they have not been mapped to the 6 CEFR levels. At test time the grader returns both a prediction of the score and the variance on the prediction. The latter can be used to reject scores in which the GP has a low confidence, but for this paper all predicted scores were retained.

To mitigate the effect of ASR errors on the grader, the GP parameters are retrained for each ASR system.

3. Off-topic Response Detection

A candidate may speak off-topic in response to a question prompt for a number of reasons. For example, they may not be able to formulate a valid answer, or not understand the question or they may "cheat" by speaking a memorised answer. If a response is not relevant to the prompt then it should receive no marks for that question, however fluent and well pronounced. The impact of ASR performance on the ability to detect off-topic responses is investigated here using a state-of-the-art attention-based relevance detector [30]. The detector derives the probability of relevance $\mathtt{P(rel}|\boldsymbol{w}^r, \boldsymbol{w}^p)$ where $\boldsymbol{w}^p$ is the prompt text and $\boldsymbol{w}^r$ is the ASR transcription of the response.

4. Experimental Results

4.1. Data and setup

Experiments to investigate the impact of ASR performance on free speaking language assessment are run on data from BULATS [16] (Section 2). The BULATS test comprises 5 sections: A. short responses; B. read aloud sentences; C-E. free speaking responses of 60 (C,D) or 20 seconds (5 parts of E) maximum length. The evaluation data set, *EVAL*, consists of 226 speakers from 6 European, Arabic and Asian L1s roughly evenly distributed across the CEFR grade range. C1 and C2 grades are merged owing to a lack of C2 speakers.

[1] System 2 outperforms a TDNN-LSTM lattice free MMI system, however it has a significant computational overhead in training and test. Further reductions in WER can be achieved through speaker adaptation and/or semi-supervised x1training but these are not considered here.

[2] There are 75 L1s in total but most have only 1 or 2 speakers.

Merged crowd-sourced transcribed [31] data from BULATS [16] is used for training and test of the two ASR systems, *System 1* and *System 2*, described in Section 2.1. Evaluation is performed on BULATS sections C-E only, *AEVAL* whereas the training data is from all sections. In both systems 9 consecutive frames of 40-D filterbank features plus their delta form the 720-D input to the bottleneck (BN) DNN. A global semi-tied covariance matrix [18] transforms the 39-D BN features which are appended to HLDA [18] projected PLP features. CMN and CVN are applied at the speaker level to yield a 78-D per frame Tandem input feature, concatenated 9$\times$ to form a 702-D input vector.

System 1 is implemented using the HTK v3.5 toolkit [32] and trained on a 108 hour, 1075 speaker, Gujarati L1 BULATS data set. The BN-DNN has a 720x1000^{4}x39x1000x6000 structure and the hybrid DNN a 702x1000^{5}x6000 structure, with global state-position triphone output targets [19]. Pre-training, cross entropy and MPE-based sequence [33] training are applied as described in [10]. A Kneser-Ney trigram LM is trained on 186k words from the System 1 training data, and interpolated with a general LM trained on Broadcast News English [34], using the SRILM toolkit [35]. A 334 hours, 8485 speaker, 28 L1s, BULATS data set is used to train System 2 which is implemented using the Kaldi toolkit (nnet) [24]. The BN-DNN and hybrid DNN-HMM have the same structure as System 1 but with 8949 tri-graphemic outputs. The LSTM has 2 hidden layers, each with 1000 memory cells and 500 recurrent projection units. Cross entropy and MPE-based sequence training are applied [33, 36]. The in-domain LM component is trained on 1.83M words from the System 2 training data.

The GP grader is trained on 994 speakers, held out from ASR, and tested on the same test set but across all sections A-E. Training and test transcriptions are generated by the ASR system under test. 33 audio and fluency related features derived from all sections are used for the baseline grader. Term frequency-inverse document frequency (TF-IDF) of 137 PoS tags are generated on the free speaking sections only [27]. The PoS tags are derived from parse trees. To determine if these trees are sufficiently robust to ASR errors to produce useful linguistic features, parse trees derived from manual and ASR transcriptions, as in Figure 2, are compared. Convolution Tree Kernels [37] are used to compute the similarities between the two parse trees for each candidate response [27].

No off-topic responses exist in the BULATS data so 10-fold cross-validation over prompts is used for testing off-topic response detection. Responses from others prompts are used as negative examples. Examples of *seen* and *unseen* prompts are observed/never observed in training, respectively. The negative responses are drawn from data held out from training. *Seen* negative responses are selected from prompts seen in training, and *unseen* from prompts not seen in training.

A subset of 1043 responses, *AEVAL1k*, [3] of the ASR evaluation data set has been manually annotated with errors and disfluencies [38]. The annotators: corrected the crowd-sourced transcriptions; performed meta-data and grammatical error correction; marked any pronunciation errors. For the pronunciation error marking, the annotators are presented with pronunciations for each word taken from Combilex [21]. They added the learner's actual pronunciation where it differed from the lexicon and added pronuniciations for OOV words [4]. Words in the corrected transcript for which the annotators provided a new pronunciation are marked as having a pronunciation error (PE). The WER corresponding to these PEs (WERPE) is then determined by time aligning the reference and ASR transcriptions. For each PE word, a word error is recorded if the reference and ASR words do not match. Words deleted by the ASR are counted as a PE with confidence score 0. Words inserted by the recogniser are ignored, as feedback is impossible in this case. Partial words, hesitations and unclear words are also ignored. The transcribers are asked to mark minimal edits to make the language grammatical and as understandable as it can be, whilst remaining faithful to the original. Based on these edits, each word in the corrected transcriptions is marked as to whether it has a grammatical error (GE). The WER of these GEs (WERGE) is computed as for WERPE but against words marked with a GE in the reference transcription.

[3] Annotation of the full set is ongoing.

[4] All the annotators reported that they had Southern Standard British English in mind as their pronunciation model [38].

4.2. Auto-marker

The recognition performance of Systems 1 and 2 are significantly different on the evaluation set as shown in Table 1. This is due to a combination of the additional AM and LM training data, covering the L1s in the test set, and the more advanced AM architecture in System 2. The assessment performance of the two systems is, however, near identical (Table 1) in terms of both Pearson Correlation Coefficient (PCC) and Mean Square Error (MSE). This indicates that the features used in the baseline automarker are able to mitigate the effects of ASR errors.

Table 1: *WER and assessment performance against expert examiners of Systems 1 and 2, and standard examiners (BULATS).*

	WER	Grader	
	(%)	PCC	MSE
System 1	47.5	0.854	11.0
System 2	30.4	0.854	11.3
BULATS	—	0.848	14.2

The lower WER System 2 does yield a noticeable improvement from System 1 in parse tree similarity to manual transcription derived trees, illustrated in Figure 3. The System 2 similarity performance is close to that of crowd-sourced transcriptions. It should therefore be sufficiently robust to generate useful PoS tags. This is seen in Table 2 where much less degradation is observed switching from the baseline to PoS only features for System 2. Combining the baseline and PoS tag features gives a PCC increase of 0.1 for System 2 but a 0.1 drop for System 1. It can be expected that more complex linguistic and content-based features would also benefit from the reduced WER of System 2.

Table 2: *Use of PoS tags for assessment.*

ASR	PCC		
	Base	PoS	Comb
System 1	0.854	0.792	0.847
System 2	0.854	0.833	0.865

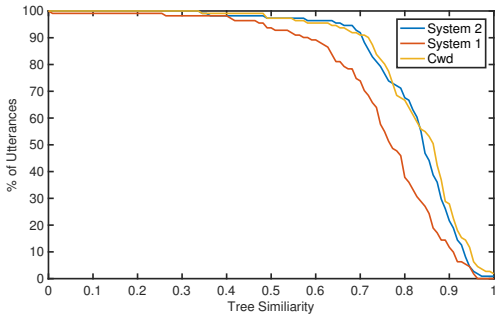


Figure 3: *Parse tree similarity to manual transcription trees of Systems 1 and 2 and crowd-sourced (Cwd) transcriptions.*

4.3. Off-topic response

Since the off-topic response requires the content of a learner's spoken response to be interpreted it would be expected to benefit from improved recognition accuracy. Table 3 shows that on all response/prompt pairs the more accurate System 2 does indeed yield better off-topic response detection with correct detection over 0.95 achieved for both types of negative responses on seen prompts. The smallest gain (0.007) is for the hardest task where both the prompts and responses are unseen in the training data.

Table 3: *Off-topic response performance of Systems 1 and 2.*

Prompts	Negative Responses from Seen Prompts		Unseen Prompts	
	Sys1	Sys2	Sys1	Sys2
Seen	0.949	0.976	0.938	0.968
Unseen	0.855	0.883	0.751	0.758

4.4. Pronunciation and grammatical errors

Pronunciation and grammar usage are key components of a learner's proficiency in English. In order to provide feedback to a learner about how what they should work on improving, an automatic system will need to detect pronunciation and grammatical errors made by the learner. As any free speaking speech recognition system will never be 100% accurate this error detection must take into account the recognition as well as language errors. Table 4 gives the WER breakdown by grade on the AEVAL1k set. As can be seen, as the grade improves the number of substitutions, deletions and insertions reduces with a corresponding reduction in WER.

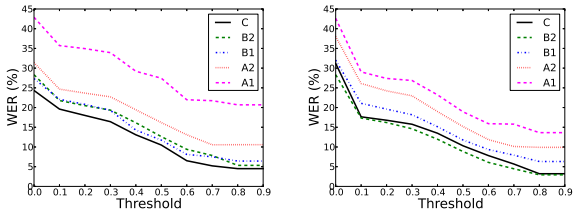
Table 4: *System 2 % substitutions (Subs), deletions (Del), insertions (Ins) and total WER by grade on AEVAL1k.*

#	A1	A2	B1	B2	C	All
Subs	18.2	14.6	13.7	13.1	11.4	13.3
Del	19.9	12.4	8.0	8.1	7.7	9.4
Ins	5.5	5.2	3.5	2.6	2.1	3.2
Total	43.7	32.2	25.1	23.7	21.2	26.0

With a confidence threshold of 0, i.e. accepting all words, the WER for words with pronunciation (WERPE) or grammatical (WERGE) errors is much higher than the WER for all words, as observed in Table 5. This is unsurprising as when a speaker mis-pronounces a word or speaks ungrammatically this makes the ASR task harder. If the confidence threshold is raised to 0.9, the proportion of recognition errors is greatly reduced. There is still a risk that the system could give incorrect feedback due to an ASR error as the WER is non-zero, however, this will occur far less often. Figure 4 shows that the WERPE and WERGE decrease with improvement in grade and increase in the confidence threshold. For WERGE there is a noticeable drop-off at threshold 0.1 as deletions are eliminated. Pronunciation feedback for the lowest grade, A1, will be most affected by ASR errors. The WERPE for speakers at level B1 is close to that of B2 and C, which is expected as at B1 pronunciations can generally be understood, with some strain due to L1 effects. More variation is seen for WERGE, where B1 speakers have more limited grammatical abilities than B2 and C speakers.

Table 5: *System 2 WER and number of errors with pronunciation (PE) and grammar errors (GE) and all words (All) without insertion errors at confidence thresholds 0 and 0.9.*

	Confidence ≥0.0		≥0.9	
	% WER	# Err	% WER	# Err
PE	29.0	993	7.3	128
GE	33.1	1611	6.3	215
All	22.7	10515	3.4	885



a) WERPE b) WERGE

Figure 4: *WER of all words with pronunciation (WERPE) and grammatical (WERGE) errors for System 2 for grades A1 to C.*

5. Conclusions

This paper considered the impact of ASR performance on assessment of free speaking tests of non-native learners of English. Comparison of 2 state-of-the-art ASR systems with a significant difference in WER showed that the baseline auto-marking system based on audio and fluency related features was unaffected by the reduction in WER due to the system having been designed to mitigate ASR errors. Improving the ASR did lead to benefits where richer information is needed, including for adding linguistically related auto-marker features derived from statistical parse trees, and detecting off-topic responses. The impact that ASR errors may have on the ability of the system to provide detailed feedback to the learner was also analysed. Although incorrect feedback on pronunciation and grammatical errors may occur from mis-interpreting ASR errors, this can be mitigated by focusing on words which the lowest WER system is most confident that it has recognised correctly.

6. References

[1] British Council, "The English Effect," Aug 2013, Research Report.

[2] K. Zechner *et al.*, "Automatic scoring of non-native spontaneous speech in tests of spoken English," *Speech Communication*, vol. 51, no. 10, pp. 883–895, 2009.

[3] AISpeech, 2012. [Online]. Available: http://bit.ly/2mMyxRX

[4] A. Metallinou and J. Cheng, "Using deep neural networks to improve proficiency assessment for children English language learners," in *Proc. of INTERSPEECH*, 2014, pp. 1468–1472.

[5] J. Butzberger *et al.*, "Spontaneous Speech Effects in Large Vocabulary Speech Recognition Applications," in *Proc. of the Workshop on Speech and Natural Language*, ser. HLT '91. Association for Computational Linguistics, 1992, pp. 339–343.

[6] Cambridge English, "Write and Improve." [Online]. Available: https://writeandimprove.com

[7] Council of Europe, *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press, 2001.

[8] Y. Qian *et al.*, "Self-Adaptive DNN for Improving Spoken Language Proficiency Assessment," in *Proc. of INTERSPEECH*, 2016, pp. 3122–3126.

[9] ——, "Bidirectional LSTM-RNN for Improving Automated Assessment of Non-native Children's Speech," in *Proc. of INTERSPEECH*, 2017, pp. 1417–1421.

[10] K. Knill *et al.*, "Use of Graphemic Lexicons for Spoken Language Assessment," in *Proc. of INTERSPEECH*, 2017, pp. 2774–2778.

[11] K. Evanini and X. Wang, "Automatic detection of plagiarized spoken responses," in *Proc. of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, 2014.

[12] S.-Y. Yoon and S. Xie, "Similarity-Based Non-Scorable Response Detection for Automated Speech Scoring," in *Proc. of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, 2014.

[13] A. Malinin *et al.*, "Off-topic response detection for spontaneous spoken english assessment," in *Proc. of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 1075–1084.

[14] M. McCarthy and R. Carter, "Spoken grammar: what is it and how can we teach it?" *ELT Journal*, vol. 49, no. 3, pp. 207–218, 1995.

[15] M. Meteer *et al.*, "Dysfluency Annotation Stylebook for the Switchboard Corpus," Linguistic Data Consortium, Tech. Rep., Feb 1995, updated June 1995 by Ann Taylor. [Online]. Available: https://catalog.ldc.upenn.edu/docs/LDC99T42/dflguide.ps

[16] "BULATS. Business Language Testing Service," Available: http://www.bulats.org/computer-based-tests/online-tests.

[17] J. Carletta *et al.*, "The AMI meeting corpus: A pre-announcement," in *Machine learning for multimodal interaction*. Springer, 2006, pp. 28–39.

[18] M. J. F. Gales, "Semi-tied covariance matrices for hidden Markov models," IEEE *Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 272–281, 1999.

[19] K.Knill *et al.*, "Language Independent and Unsupervised Acoustic Models for Speech Recognition and Keyword Spotting," in *Proc. of INTERSPEECH*, 2014.

[20] G. Evermann and P. Woodland, "Large vocabulary decoding and confidence estimation using word posterior probabilities," in *Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2000.

[21] K. Richmond, R. A. J. Clark, and S. Fitt, "On generating Combilex pronunciations via morphological analysis," in *Proc. of INTERSPEECH*, 2010, pp. 1974–1977.

[22] M. Bisani and H. Ney, "Joint-Sequence Models for Grapheme-to-Phoneme Conversion," *Speech Communication*, vol. 50, pp. 434–451, May 2008.

[23] H. Wang *et al.*, "Joint decoding of tandem and hybrid systems for improved keyword spotting on low resource languages," in *Proc. of INTERSPEECH*, 2015, pp. 3660–3664.

[24] D. Povey *et al.*, "The Kaldi Speech Recognition Toolkit," in *Proc. of the Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2011.

[25] D. Higgins *et al.*, "A three-stage approach to the automated scoring of spontaneous spoken responses," *Computer Speech and Language*, vol. 25, no. 2, pp. 282–306, 2011.

[26] R. van Dalen, K. Knill, and M. Gales, "Automatically Grading Learners' English Using a Gaussian Process," in *Proc. of Workshop on Speech and Language Technology for Education (SLaTE)*, 2015.

[27] Y. Wang *et al.*, "Towards Automatic Assessment of Spontaneous Spoken English," 2017, manuscript submitted for publication in Speech Communication.

[28] H. Yannakoudakis, T. Briscoe, and B. Medlock, "A new dataset and method for automatically grading ESOL texts," in *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 2011, pp. 180–189.

[29] T. Briscoe, J. Carroll, and R. Watson, "The second release of the RASP system," in *Proc. of the COLING/ACL on Interactive presentation sessions*. Association for Computational Linguistics, 2006, pp. 77–80.

[30] A. Malinin *et al.*, "An attention based model for off-topic spontaneous spoken response detection: An Initial Study," in *Proc. of Workshop on Speech and Language Technology for Education (SLaTE)*, 2017.

[31] R. C. van Dalen, K. M. Knill, P. Tsiakoulis, and M. J. F. Gales, "Improving multiple-crowd-sourced transcriptions using a speech recogniser," in *Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Apr 2015.

[32] S. Young *et al.*, *The* HTK *book (for* HTK *version 3.5)*. University of Cambridge, 2015. [Online]. Available: http://htk.eng.cam.ac.uk

[33] K. Veselỳ *et al.*, "Sequence-discriminative Training of Deep Neural Networks," in *Proc. of INTERSPEECH*, 2013, pp. 2345–2349.

[34] J. S. Garofolo, J. G. Fiscus, and W. M. Fisher, "Design and Preparation of the 1996 Hub-4 Broadcast News Benchmark Test Corpora," in *Proc. of DARPA Speech Recognition Workshop*, 1997.

[35] A. Stolcke, "SRILM - An Extensible Language Modeling Toolkit," in *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, 2002, pp. 901–904.

[36] H. Sak *et al.*, "Sequence Discriminative Distributed Training of Long Short-Term Memory Recurrent Neural Networks," in *Proc. of INTERSPEECH*, 2014, pp. 1209–1213.

[37] M. Collins and N. Duffy, "Convolution kernels for natural language," in *Proc. of the Conference on Neural Information Processing Systems (NIPS)*, 2001, pp. 625–632.

[38] A. Caines, D. Nicholls, and P. Buttery, "Annotating errors and disfluencies in transcriptions of speech," University of Cambridge, Computer Laboratory, Tech. Rep. UCAM-CL-TR-915, Dec. 2017. [Online]. Available: http://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-915.pdf